Data Misakh Team

Classify cancer types using gene expressions

Health and medicine

Arash Azhand
Danial Hadizadeh
1404/02/22

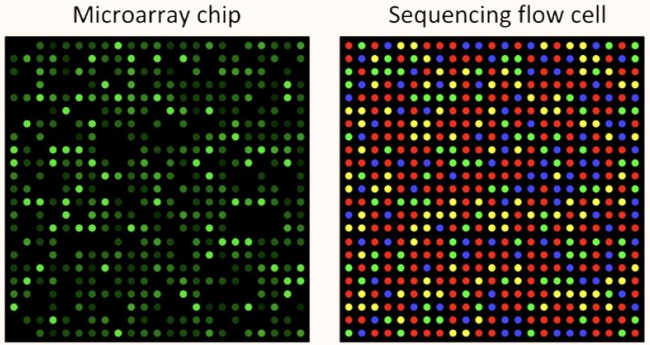
Table of contents

Introduces the goal of **classifying cancer** types using **gene expression** data, and highlights how understanding gene activity can support early diagnosis and personalized treatments.

01 Introduction

Introduction - Understanding the problem

- **What are gene expressions?**
- **How do we measure them?**
  - Using microarrays



- **Why they matters in cancer?**
  - Some genes are **overexpressed** -> those promoting cell growth
  - Some are **underexpressed** -> those suppress tumors

Describes the structure and **challenges of the dataset**, including the high number of gene features, limited samples, and the presence of six distinct cancer types.

02
Dataset Description

About dataset

54,675 gene expression values for each **151 samples** (**normalized**)

- Each sample = 1 instance
- Each gene = 1 feature

6 classes: 1.Basal 2.HER 3.luminal_B 4.luminal_A 5.cell_line 6.normal

- Why we expect it to be a **hard dataset**?

  This is a **high-dimensional**, **low-sample**, **multi-class** classification problem!

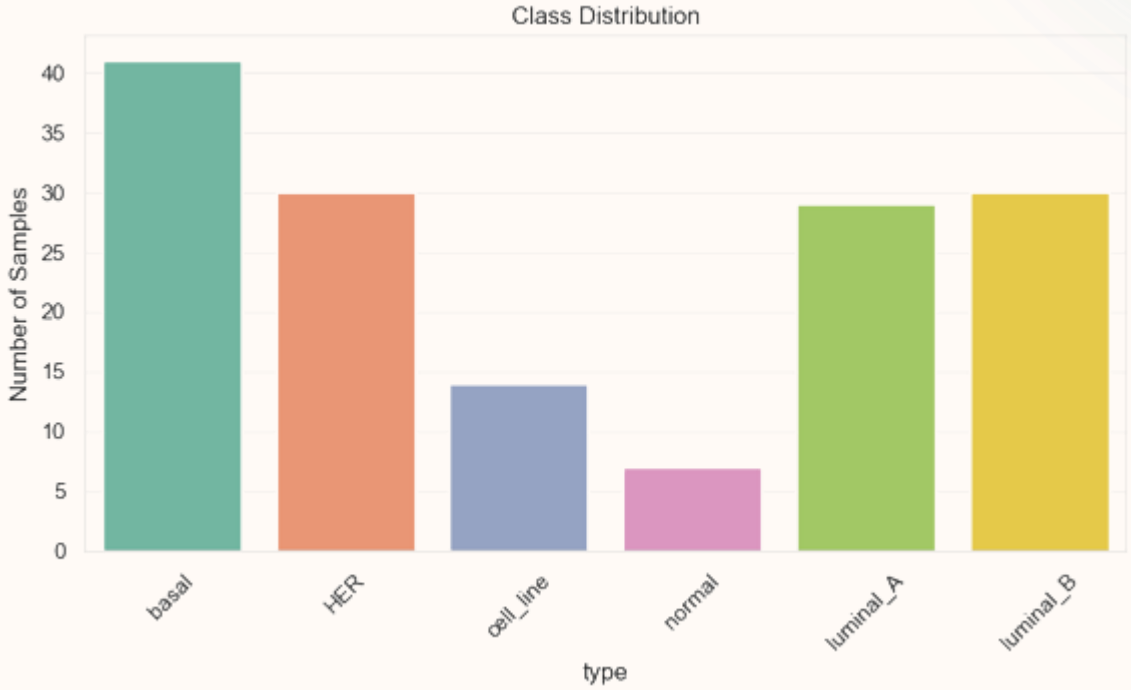
  → Curse of dimensionality

Presents initial observations about class distribution and gene expression variability, helping to **identify potential issues** like class imbalance and noise in the data.

03
Data Exploration

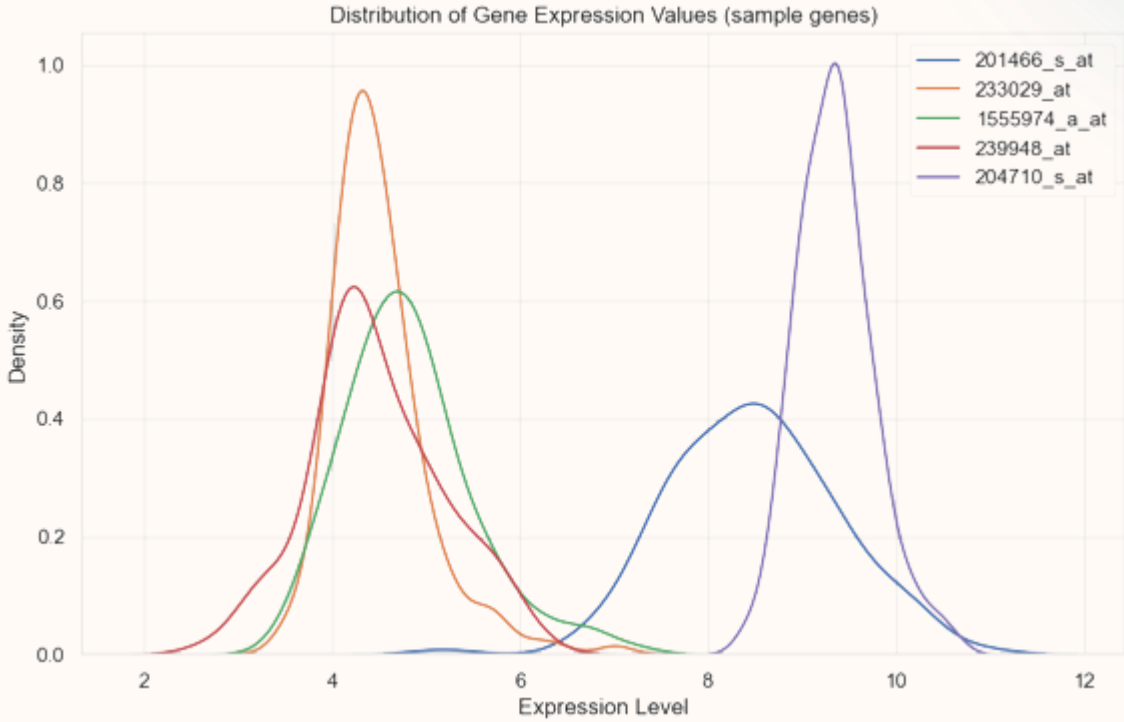
Data Exploration (Class Distribution)

- We also have class imbalance too!



Data Exploration (Gene Distribution)

- Distribution of 5 random genes:



Explains how uninformative genes were removed to **reduce dimensionality**, using statistical methods like variance filtering and ANOVA to retain the most relevant features.

04

Data Preprocessing & Feature Selection

Variance thresholding

To start reducing dimensionality, we removed genes with very low variance across samples, as they are unlikely to help distinguish cancer types.

- **Original shape:** (151, 54675)
- **Threshold:** variance < 0.01
- **After filtering:** (151, 54605) → **70 genes dropped**

This simple step removes non-informative features and slightly reduces noise before deeper feature selection.

ANOVA F-test

We then used ANOVA F-test to rank genes based on how well they separate cancer types.

- **Idea:** Low F-score → gene is similar across all classes → not useful
- **Formula:** F = Between-class variance / Within-class variance
- Top **1000 genes** selected → New shape: **(151, 1000)**

This step keeps only the most class-informative genes and greatly improves learning efficiency.

Why 1000 genes? what's the difference with var. threshold?

Covers the training and evaluation of multiple classification models using cross-validation, aiming to **find the best approach** for handling the complex dataset.

05

Machine Learning Models

Comparing models

We used 3 models along with stratified 5-fold cross validation to generalize our result:

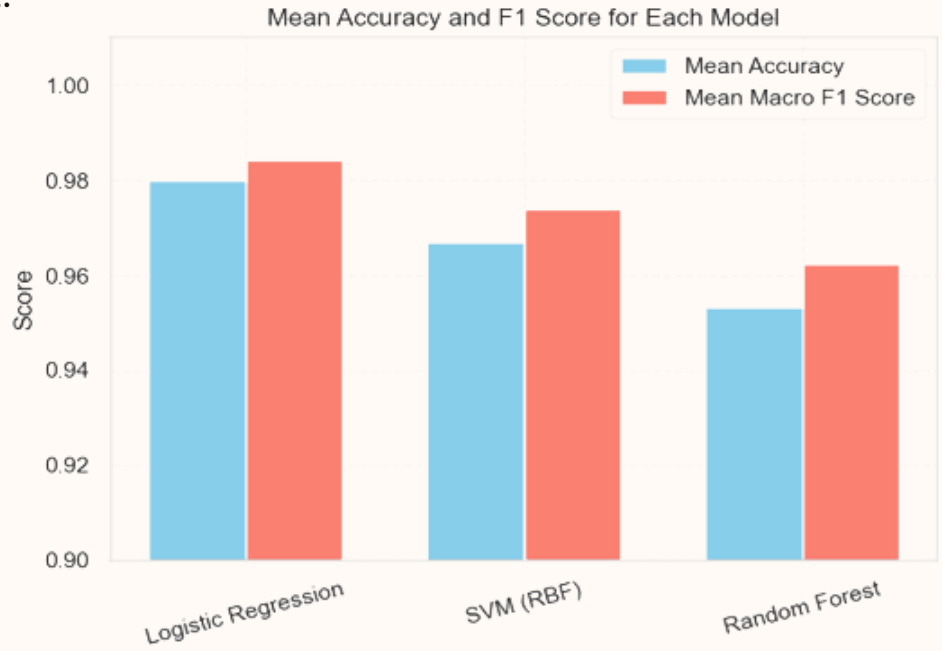
01. **Logistic Regression**

02. **SVM (RBF)**

03. **Random Forest**

These accuracies are

Surprisingly high!!



Highlights the **performance difference** between models trained on raw vs. reduced data, demonstrating the impact of proper feature selection on accuracy and stability.

06

Comparison with Raw Dataset

Compare with Original Dataset (without Feature Selection)

Model	Accuracy (Full)	Accuracy (Reduced)	Δ Accuracy	Δ F1 Macro
Logistic Regression	89.4%	98.0%	▲ +8.6%	▲ +9.8%
SVM (RBF)	93.3%	96.7%	▲ +3.4%	▲ +5.6%
Random Forest	95.4%	95.3%	▼ ~same	▲ +3.0%

So these feature selection did really help models to **perform better!**

Discusses why the models performed well despite data complexity, and uses PCA visualizations to show how cancer types separate in lower-dimensional space.

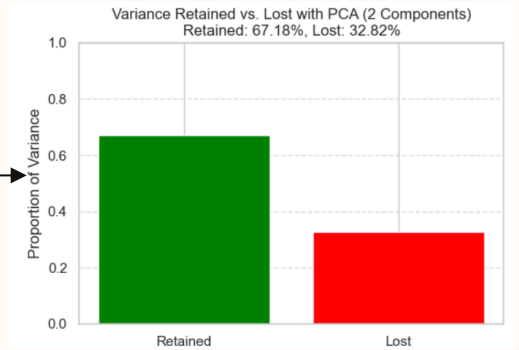
07

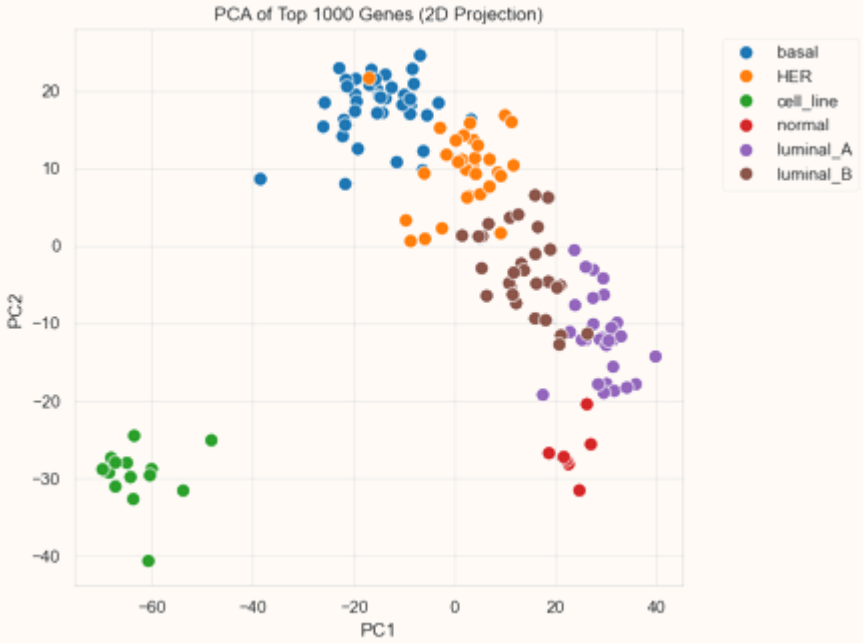
Discussion & Visualization

Discussion

- **Why are we even getting this good accuracy on this hard dataset??**
- Curated dataset
- We used PCA to project the data into 2D space
- Different classes are almost **linearly Separable** with PCA!

Information loss →





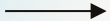
08
Conclusion

Conclusion

- ✓ High-dimensional gene expression data
- ✓ Importance of feature selection (Variance Threshold + ANOVA F-test)
- ✓ Logistic Regression sensitive to irrelevant features
- ✓ Random Forest performed best overall
- ✓ PCA showed clear class separability

Resources

- **Articles & Papers:**
  - [1] CuMiDa: An Extensively Curated Microarray Database for Benchmarking and Testing of Machine Learning Approaches in Cancer Research
  - [2] Neuroevolution as a tool for microarray gene expression pattern identification in cancer research



Do you have any questions?

a.azhand@ec.iut.ac.ir
d.hadizadeh@ec.iut.ac.ir

Thanks!